ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(13), DOI: 10.17485/ijst/2016/v9i13/85481, April 2016* ISSN (Online) : 0974-5645

Survey on Web Mining Techniques and Challenges of E-commerce in Online Social Networks

K. N. Asha and R. Rajkumar

School of Computing Science and Engineering, Vellore Institute of Technology, Vellore – 632014, Tamil Nadu, India; cuashin@gmail.com, vitrajkumar@gmail.com

Abstract

Objectives: In recent years, we have tremendous growth of users in Online Social Networks (OSN) such as Facebook, Google+, twitter etc. This becomes major reason for enabling web as largest market defining E-Commerce. Many companies use OSN as their sales channel as they can reduce operating cost for managing orders significantly compared with traditional channels. Also viral marketing is very popular in OSN. But the major drawback of this channel is if the company doesn't satisfy the customer, then the same OSN can rapidly propagate a bad reputation of the company affecting the company's business. Hence for customers it has become very important to identify and filter dishonest recommenders. So it becomes very important to recommend right items to right customers. In this survey, we aim to give a comprehensive review of research related to E-Commerce in OSN. **Methods:** First, we discuss web usage mining techniques for better shopping websites to satisfy customers. Also we discuss web mining techniques to find dishonest recommenders in OSN. **Findings/ Improvements:** Our survey explores the existing research highlights and also presents various researches in these topics. Also, we propose recommendation system which uses Semantic Web Mining process integrated with domain ontology which can be used to extract interesting patterns from, complex and heterogeneous data.

Keywords: E-Commerce, OSN, Recommender System, Spammer Detection, Web Mining Technologies

1. Introduction

Today, internet has become the best medium of communication, which allows information exchange among users connected globally. The 60% of ACTIVE internet users have at least one profile in one of the popular social networks, such as Facebook, twitter etc. This becomes major reason for enabling web as largest market defining E-Commerce. Many companies use OSN as their sales channel as they can reduce operating cost for managing orders significantly compared with traditional channels. As OSN are in prime focus for mining the social/ customer data web/data mining should be integrated with E-commerce applications to improve the performance of E-commerce companies[1,2,3]. Also, OSN users share the information about purchased items with their friends, or they may seek recommendations about the item purchased from their friends and even they may recommend their friends to buy an item. Thus one influences the other to do the purchase which spreads quickly in OSN thus increasing company's sales. This type of influence/advertisement is referred to as social influence or viral marketing.[4,5] Users in social networks and their relations with other user in the social network should be analysed using web mining technique so as to avoid bad mouth in viral marketing.

2. Web Usage Mining for Shopping Websites

2.1 Motivation and Challenges

As OSN has become an important platform for online shopping. Online Recommendation is being used by many shopping websites. Recommendation should be done based on the user's interest on items. Hence recommender system should provide the currents needs of user. Thus this system should be based on user's navigational patterns[6,7].

Using Web usage mining there are many different ways through which a recommender system can be created,

* *Author for correspondence*

[6,8-11]Some algorithms used in general are:

- Collaborative filtering
- Content Based Filtering

2.1.1 Collaborative Filtering

The recommender system uses data analysis methods which help the users to find the items they want to buy at E-Commerce websites by generating a predicted likeliness score or a list of top-N items to be recommended for a given user. The item can be recommended using various methods. Item can be recommended based on user behaviour, with the past purchasing pattern of user as a predictor of future purchase. The basic idea behind Collaborative Filtering algorithm is to give recommendations based on the views of like-minded users. The views of users can be found explicitly from the users or by means of some implicit procedures. The item-based method collects the set of items that are rated by the target user and calculates how similar they are to the items that are targeted. Once the most similar items are found, the prediction is then calculated by taking a weighted average of the ratings given by target user on these similar items.

2.1.2 Content based Filtering

The basic concept of content-based filtering system is to select items based on the association between the item's content and the items preferred by users as opposed to a collaborative filtering based system which selects items based on the association between users with similar likings. It recommends the user by comparing a user profile with each document's content in the collection. Document's content can be characterised with a set of terms. By running through a many parsing steps, these terms are mined from documents. The user profile is characterised with the similar terms and constructed by investigating the document's content that the user found interesting. The documents that are of user's interests can be found by using either explicit feedback which involves the user to evaluate examined documents on a scale or implicit feedback in which the user's interests are inferred by observing the user's actions, although it is more convenient for the user but is more difficult to implement.The existing system[5] have combined algorithms and achieved efficiency such as high speed, minimum memory usage. This System works in 5 phases:

- **Phase 1: Sort Phase**

The database (D) is sorted, with the major key being customer-id and the minor key as transaction-time. This step implicitly converts the original transaction database into a database of customer sequences.

- **Phase 2: Large Item-set Phase**

Here the set of all L-item sets are found.

- **Phase 3: Transformation Phase**

In a converted customer sequence, every transaction is replaced by the set of all L-item sets contained in that transaction. If a transaction does not contain any l-item set, it is not retained in the transformed sequence else if sequence does not contain any l-item set, this sequence is removed from the converted database. However, it still gives information about the total number of customers.

- **Phase 4: Sequence Phase**

This phase uses the set of item sets to find the preferred sequences. The two algorithms of this phase are, Apriori some and Apriori All. They have comparable performance; although Apriori some performs a slight better when the minimum number of customers that must support a sequential pattern is low. The main advantage of *Apriori Some* over *Apriori All* is that it avoids counting many non-maximal sequences.

The future work concentrates on the recommender systems that have been widely used in several OSN. However, there is always a scope to improve the quality of recommendations and the user satisfaction with the recommendation systems.

3. Dishonest Recommenders in OSN

3.1 Motivation and Challenges

OSN users share the information about purchased items with their friends, or they may seek recommendations about the item purchased from their friends and even they may recommend their friends to buy an item. Thus one influences the other to do the purchase which spreads quickly in OSN thus increasing company's sales. However,

there are some dishonest users in the same OSN who give misleading recommendation to their neighbours which opens door for malicious activities.

3.2 Existing Solution and Discussion

3.2.1 Basic Detection Algorithm[13]

This algorithm if fully distributed. The users here can independently perform to recognize dishonest user's among their neighbors. With no loss of generality, the only focus is on one specific user, say user *I, called the detector*. The algorithm is from the viewpoint of user *I* in detecting dishonest neighbors. The algorithm represents product as a trustworthy or untrustworthy product if the detector considers it to be trustworthy or untrustworthy.

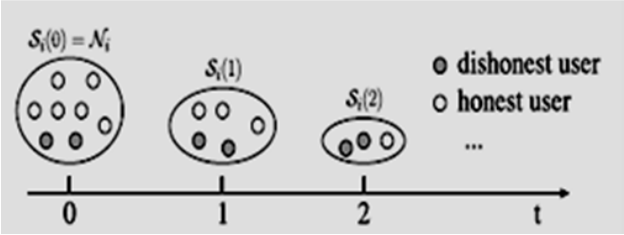

Figure 1. Detection Framework to minimize the suspicious set.

To illustrate the performance of the detection algorithm, three performance measures are defined:

- *Probability of false negative* which is denoted as *Pf n(t)*,
- *Probability of false positive* which is denoted as *Pf p(t)*, and
- The *number of iterations needed to minimize the suspicious set until it contains only dishonest users*, which is denoted by a random variable *R*.

Thus *Pfn(t)* is denoted by,

$$\text{Pf n(t)} = \frac{\text{total number of dishonest neighbors of i not in Si(t)}}{\text{total number of dishonest neighbors of detector i}}$$

The detection framework given here can be observed as a valuable framework to sustain the feasibility of viral marketing in OSNs. Further, the work can be enhanced by implementing more trust algorithms to build a framework to perform on large-scale experiments which uses large data sets from OSN.

4. Spammer Detection by Comparing user Behaviors

4.1 Motivation and Challenges

Today, OSN users are receiving requests and messages from unwanted friend, in their OSN accounts, which refer to the introduction of spammers in social networks. Advertising, Phishing, Malware, etc. contained in spam can always a key to user's privacy disclosure. As there is a great impact of spammers in OSNs, it is very important to detect them.

4.2 Existing Solution and Discussion[14]

Analyze the user's data for detecting spammers is based on the user behaviour in OSN. User behaviors in social networks are categorized into: *relation-related* behaviors *and tweet-related* behaviors[14].

4.2.1 Information used for Spammer Detection

To analyze user behaviour, a user is selected in random; then crawled the first two hops following it and then crawled their relationship to form a complete sub-graph. Again the crawling into the messages they posted is done in this sub-graph in order to get further behavior details. Here, the users sending undesirable following request or advertisements in OSNs are labeled as spammers. This information set is further used to conduct behavior analysis.

4.2.2 Relation-related Behaviour: Relation Creation

From the study it is observed that spammers usually follow many users than the normal users in OSNs. This is mainly because, in OSN, normal users spend less time and hence follow less number of users, resulting in fewer bi-followers. The bi-followers always indicate true friendship. Whereas, the spammers spend more time on OSN as they are meant to spread advertisements or malware and hence they have many followers but less bi-followers. Thus, the information about friend-bi-follower ratio can be used to detect spammers in OSN.

4.2.3 Message-related Behaviour: User Activeness

User activity is defined as the amount of Messages users generate per month and present its distribution. As the

spammers are meant to spread advertisements or malware, they produce more messages than the normal user. Also the spammers produce messages more frequently and hence they are more active than the normal users.

4.2.4 Message-related Behaviour: User Interaction

Users interactions are defined as the exchange of messages between users including *reply* and *repost in OSN*. From the studies it is observed that spammers have lesser friends and hence they will have lesser replies because, replies are interactions that are to be expected between close friends. Also, as they have lesser replies, replies cannot be used to spread the advertisements or malwares and hence, spammers always select to repost a message instead of giving a reply.

4.2.5 Message-related Behaviour: Message Content

With further investigation it is found that there is are lationship between spammer detection and the message content features, including *message length* and *hyperlink ratio*. Lengthier messages include more information, whereas URLs always a main link for shopping, malwares or phishing etc. Thus, as spammers are meant to spread advertisements or malware spammers etc. publish lengthier messages so as to provide more information about the advertisements. The spammers also desire to create a message with a hyperlink in it so that when users click the hyperlink spammers achieve their malicious activity. Due to dynamic patterns of spammers, it becomes very important to discover attributes to separate spammers from normal users automatically.

5. Proposed Work

To build a good recommendation system we propose a system which uses Semantic Web Mining process integrated with domain ontology which can be used to extract interesting patterns from, complex and heterogeneous data.

As the architecture of Semantic Web[15,16] explains the various advantages layers including trustworthiness of information, this can be used for building a trustworthy recommender system which avoids dishonest users.

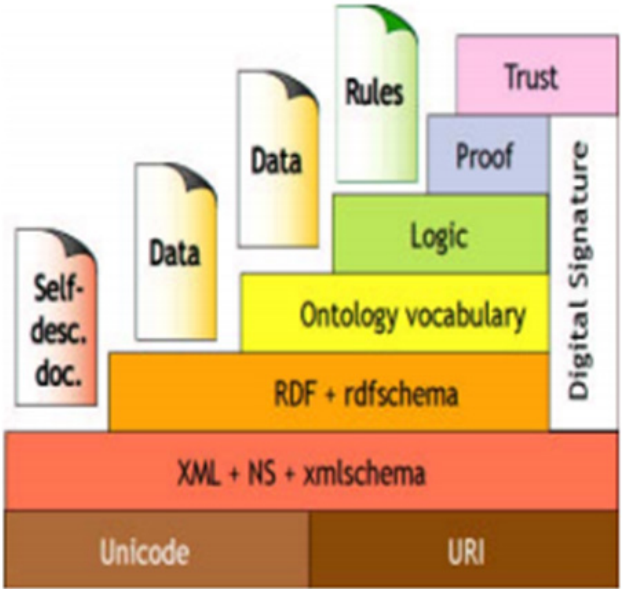

Figure 2. Semantic web architecture.

The system proposed here, uses ontology based representation of the items and user-profiles. The system consists of 5 phases shown below:

- **Phase 1: Enumerate important terms in ontology**
  It is useful to make a set of all desired terms.

- **Phase 2: Collect URL of the web page visited**
  We find the set of all item sets consisting of navigated URLs by user.

- **Phase 3: Term Extraction from web page URL**
  We obtain the term of the domain that we want to create the ontology. We extract terms using term extraction techniques. The output will be a list of concepts.

- **Phase 4: Build Semantic Network**
  A Semantic network is a propositional knowledge structure consisting of set of nodes that are connected by links labeled by the relation between each pair of connected nodes.

- **Phase 5: Sequence Phase**
  Keep Searching for the sequential patterns that matches the terms until frequent sequences of the set are not found.

6. Conclusion

Many companies are redefining their business strategies to improve the business output. In this survey, we made an effort to study Web mining techniques and challenges in three different aspects: modelling a shopping websites, understanding user behaviour to detect dishonest users in OSN, spammer detection in OSN. We review existing schemes and also provide potential future direction focusing mainly on the OSN's.

7. References

1. Siddiqui AT, Aljahdali S. Web mining techniques in e-commerce applications. International Journal of Computer Applications (0975 – 8887). 2013 May; 69(8):39–43.
2. Purandare P. Web mining: A key to improve business on web. IADIS European Conference Data Mining. 2008.
3. Satish B, Sunil P. Study and evaluation of user's behaviour in e-Commerce using data mining. Research Journal of Recent Sciences. 2012; 1:375–87.
4. Dhawan S, Singh K, Khanchi V. Critical analysis of social networks with web data mining. IJITKM Special Issue (ICFTEM-2014). 2014 May. p. 107–11.
5. M. Vedanayaki. A Study of data mining and social network analysis. Indian Journal of Science and Technology, 2014 Nov; 7(S7):185–7.
6. Iyer N, Dcunha A, Desai A, Jain K. Survey on online recommendation using web usage mining. International Journal of Computer Science and Information Technologies. 2015; 6(2):1465–7.
7. Lim M, Byunand H, Kim J. A web usage mining for modeling buying behavior at a web store using network analysis. Indian Journal of Science and Technology. 2015 Oct; 8(25).
8. Jafari M, Sabzchi FS, Rani AJ. Applying web usage mining techniques to design effective web recommendation systems: A case study. ACSIJ Advances in Computer Science: An International Journal. 2014 Mar; 3(2(8)):78–90.
9. Ya L. An Analysis of Web Mining-based Recommender Systems for E-commerce. The 2nd International Conference on Computer Application and System Modeling; Paris: France. Atlantis Press; 2012.
10. Babu KG, Komali A, Mythry V, Ratnam ASKS. Web Mining using Semantic Data Mining Techniques. International Journal of Soft Computing and Engineering. 2012 July; 2(3):168–71. ISSN: 2231-2307.
11. Xu G, Zhang Y, Li L. Web mining and social networking techniques and applications. New York, Dordrecht Heidelberg London: Springer; 2011. Available from: 10.1007/978-1-4419-7735-9
12. Kathirvel P. A survey on online shopping recommendation based on customer transactions. International Journal of Science, Engineering and Technology Research. 2015 Mar; 4(3):564–66.
13. Li Y, Lui JCS. Friends or foes: Distributed and randomized algorithms to determine dishonest recommenders in online social networks. IEEE Transactions on Information Forensics and Security. 2014 Oct; 9(10):1695–1707.
14. Chen Z, Yang J, Wang JH. A Cascading framework for uncovering spammers in social networks. Networking Conference; Trondheim. 2014 June; 2(4):1–9.
15. Dhawan S, Singh K, Khanchi V. Critical analysis of social networks with web data mining IJITKM Special Issue (ICFTEM-2014). 2014 May. p. 107–11.
16. Ting IH, Wu HJ. Web mining applications in e-commerce and e-services. Studies in Computational Intelligence. Springer-Verlag Berlin Heidelberg; 2009. p. 172.